El-WOZ: a client-server wizard-of-oz interface

Thomas Pellegrini[1,2], **Vahid Hedayati**[2], **Angela Costa**[2,3]
[1] IRIT - Université Paul Sabatier
118 route de Narbonnes, F-31062 Toulouse Cedex 9 France.
[2]INESC-ID
Rua Alves Redol, 9
1000-029 Lisbon Portugal.
[3] CLUNL
Avenida de Berna 26-C
1069 - 61 Lisbon Portugal.
pellegri@irit.fr, vahidhdyt@gmail.com, angela@l2f.inesc-id.pt

Abstract

In this paper, we present a speech recording interface developed in the context of a project on automatic speech recognition for elderly native speakers of European Portuguese. In order to collect spontaneous speech in a situation of interaction with a machine, this interface was designed as a Wizard-of-Oz (WOZ) plateform. In this setup, users interact with a fake automated dialog system controled by a human wizard. It was implemented as a client-server application and the subjects interact with a talking head. The human wizard chooses pre-defined questions or sentences in a graphical user interface, which are then synthesized and spoken aloud by the avatar on the client side. A small spontaneous speech corpus was collected in a daily center. Eight speakers between 75 and 90 years old were recorded. They appreciated the interface and felt at ease with the avatar. Manual orthographic transcriptions were created for the total of about 45 minutes of speech.

Keywords: speech recording interface, wizard-of-oz, European Portuguese elderly speech

1. Introduction

The work described in this paper was done in the framework of the "AVoz" project, a national Portuguese project about both language and acoustic models for Automatic Speech Recognition (ASR) for the elderly[1]. Due to both cognitive and physiological age-related changes, elderly speech shows specific characteristics that make its processing significantly harder when using models built using speech from younger people. In particular, automatically recognizing the speech of older people is known to be challenging compared with automatically recognizing the speech of younger people, with performance decreases of around 9-12% absolute (Baba et al., 2004; Vipperla et al., 2008). In a previous study (Pellegrini et al., 2012), we used a large read speech corpus in European Portuguese (EP) to measure significant performance differences among age groups ranging from 60- to 90-year-old speakers. An increase of 41% relative (11.9% absolute) in word error rate was observed between 60-65-year-old and 81-86-year-old speakers.

This difference in performance can be explained by the fact that the statistical models (the phone-like acoustic models) are usually trained with young adult speech, for applications such as broadcast news transcription. In order to narrow this ASR performance gap, it is necessary to collect elderly speech and adapt the acoustic models to elderly speakers (above 70 years old). For this purpose, we developed a Wizard-of-Oz (WOZ) interface. In a WOZ setup, users interact with a human wizard but they think they are interacting with an automated dialog system (Dahlbaeck et al., 1993). This method is an invaluable tool for investigating different design options of spoken dialog systems, without having to implement these systems. WOZ fake dialog systems are often used to collect atypical speech data such as elderly speech. It is the case of the MATCH and the JASMIN-CGN corpora for example. For the MATCH corpus, the simulated dialog task consisted in making an appointment with a physiotherapist. For our experiment, the dialog consisted in asking questions related to the general health conditions of the speaker.

In this paper, we describe our tool called El-WOZ, and also the small European Portuguese speech corpus that was collected with the interface. Elderly speakers interacted and were recorded with the interface. Both the corpus and parts of the wizard interface will be made available under a METASHARE license. The software was developed to collect elderly speech but it can be used for any type of speech data collection. For our speech collection, subjects interacted with a talking head. The human wizard chooses pre-defined questions or sentences in a graphical user interface, which are then synthesized and spoken aloud by the avatar on the client side. The subject interacts with the avatar believed to be autonomous. The main interesting feature of El-WOZ compared to other available WOZ software remains in the fact that it is implemented as a client-server application, which allows the wizard to control the avatar and collect speech remotely via an Internet connection.

The paper is organized as follows: Section 2. is a technical description of the interface implementation, Section 3. reports on a first use of the interface and on the corpus that was collected. Finally, Section 4. reports on preliminary ASR results on this corpus.

[1]http://avoz.l2f.inesc-id.pt

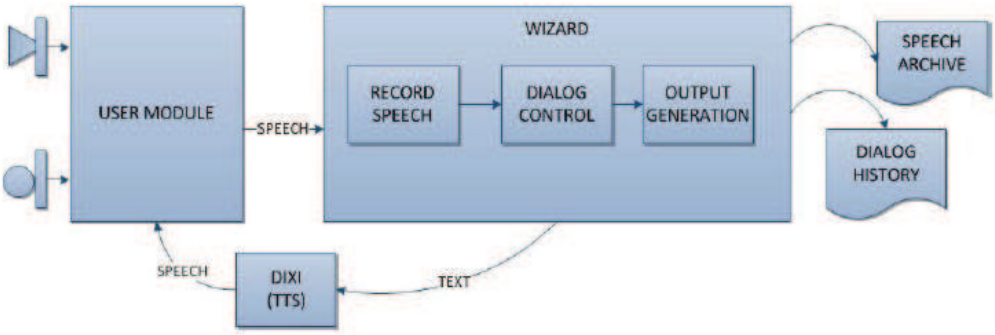

Figure 1: General architecture of EL-WOZ

2. Interface description

The standard technique often used to collect "atypical" speech data makes use of Wizard-of-Oz environments (WOZ). In a WOZ setup, users interact with a human wizard but they think they are interacting with an automated dialog system (Dahlbaeck et al., 1993). This method is an invaluable tool for investigating different design options of spoken dialog systems (SDS), without having to implement these systems.

Figure 1 shows the global architecture of our WOZ system. It is composed of two main parts: the *User Module* on the client side and the *Wizard* interface on the server side. When both applications have started and the connection has been established, the experiment can be initiated. The wizard user on the server side may select one of the sentences among a set of predefined sentences. The sentence is then sent to a text-to-speech module and then to the user module.

2.1. The user module

The user module (client-side) is a simple HTML page that displays an avatar (we used a talking head for our recordings). Below the avatar frame lies a large speak-to-talk button. The page is shown in Figure 2. The button was initially configured to be clicked each time the subject answers to the avatar. In the final version of this module, we changed this setting. The button needs to be clicked only once at the beginning of the interaction. In the first setting, each time the subject clicked on the button, a new audio file was recorded on the server side. In the second case, a single audio file is recorded from the beginning to the end of the interaction. This last option appeared to alleviate the interaction, especially with elderly subjects.

2.2. The Wizard module

The Wizard module includes three elements: Record Speech (RS), Dialog Control (DC), and Output Generation (OG). After receiving speech on the server side, the RS module converts it into a WAV file and archives it in a speech archive folder. The user name entered before the interaction and the date and time are used to automatically use name the audio files.

The DC object plays the managing and controlling role of the system. It is responsible for replying to the user by giving a suitable answer or by asking a question to go on in

Figure 2: The User Module: a simple Web page with a talking head.

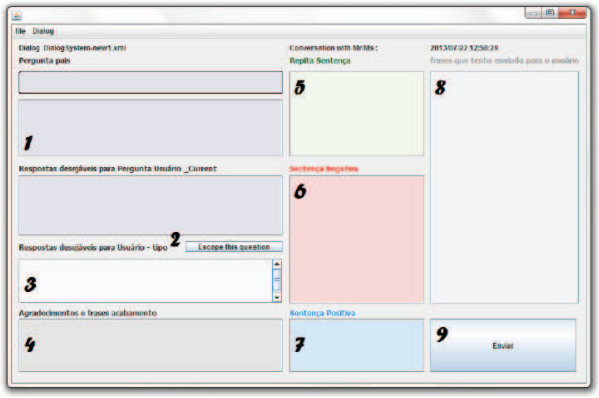

Figure 3: The Dialog Control interface.

the dialog. The wizard user controls this object. Figure 3 shows the DC interface. Multiple boxes correspond to different kinds of questions or answers to be spoken by the talking head. The box number 1 located on the left-hand side of the figure corresponds to questions of the predefined scenario i.e. the main questions to be asked to the subject.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<DialogSystem>
<Questions>

<Question  parent="0" >Olá. Sou a Joana e sou a enfermeira virtual.</Question>
<Question  parent="0" >Bom dia. Sou a Joana e sou a enfermeira virtual.</Question>
<Question  parent="0" >Boa tarde. Sou a Joana e sou a enfermeira virtual.</Question>
<Question  parent="2">Para começar, como é que se chama?</Question>
<Question  parent="2">Como é que a Senhora se chama?</Question>
<Question  parent="2">Como é que o Senhor se chama?</Question>
<Question  parent="2">Qual é o seu nome?</Question>

</Questions>
</DialogSystem>
```

Figure 4: An excerpt of a XML file used to specify a set of questions and sentences in European Portuguese.

Button number 2 allows escaping the current question without sending it. In the box number 3 the wizard can type any sentence if the predefined questions or answers are not satisfactory at a given moment of the interaction. The box number 4 is the set of thanking sentence for finishing the conversation. The box number 5 is the set of repeat request sentences, the box number 6 corresponds to disagreement, the box number 7 contains agreement sentences and the box number 8 on the right is the dialog history. Any sentence or phrase is sent to the synthesizer by clicking on the button number 9.

The OG module is responsible for sending the text of the question or answer of the avatar to our external text-to-speech software named DIXI, which runs as a Web service (Paulo et al., 2008). This engine uses the standard concatenative unit selection synthesis. All the synthetic sentences used the same female voice ("Joana") since the talking head was a female character. The synthesized speech is sent to the user module. The other task of this module consists in sending a copy of all the questions to the history dialog folder to keep track of the interaction.

Dialogs can be created within the application, which allows the creation of a new dialog from scratch, by manually adding questions. Pre-defined question and answer XML files can also be imported directly. Five distinct XML files are required: one with the main questions, one with positive answers and back-channel, one with negative answers, one with phrases asking the subject to repeat, and one with phrases to end the dialog and thank the subject. An excerpt of the main question XML file is shown in Figure 4.

3. Collecting elderly speech

El-WOZ was developed in order to record spoken interactions between elderly speakers and a simulated dialog system. We conducted several recording sessions in a day center for the elderly in Lisbon. People above 80 years old, mainly women, frequent the center on a daily basis. Although they were not used to interact with computers, they were very cooperative and eager to participate to the experiment. A set of 28 questions about general health care was used. The avatar asked questions about their sleep, their blood pressure, their appetite, but also about social aspects. The health-related questions, such as *P*osso lhe perguntar se é fumador (fumadora)?, "May I ask you if you smoke tobacco?" could be simply answered by yes or no, but most of the subjects developed their answers. The social-related questions, such as *V*ê os seus amigos muitas vezes? "do you spend much time with your friends?", also were a pretext to collect longer answers in a pretty spontaneous speech style. We tried to formulate the questions and answers in a very polite way that is known to be particularly appreciated by the elderly. In fact, most of them ended the conversation by thanking the avatar and by calling it "nurse".

It was also very interesting to see that they interacted with the avatar as if the avatar was understanding everything they said. They also had no idea that a real dialog system would make errors, even if we simulated mis-understandings.

During the first session, the subject was asked to click the "push-to-talk" button between each question and before answering to the avatar. It revealed an impossible task since they were forgetting to do it and they answered as soon as they understood the question, even before the avatar had finished his sentence. Thus, we changed the button function, which had to be clicked only once at the beginning of an interaction.

So far 11 speakers were recorded. Each speaker session lasted about 15 minutes. A headset phone was used. In total, about 4.5K words were manually transcribed for the 11 interactions that were recorded so far, corresponding to about 40 minutes of speech after removing silent segments. Table 1 shows the corresponding numbers. The final speech corpus, named "Centro de Dia" corpus will be made available at our Metashare Website[2], along with the manual orthographic transcriptions.

Gender	**# Spk**	**Duration** (min)	**# Word** **Types**	**Tokens**
Female	10	41	358	4401
Male	1	2	23	93

Table 1: Duration and number of words of the "Centro de Dia" speech corpus

[2]http://metanet4u.l2f.inesc-id.pt/

4. Preliminary ASR experiments

The ASR system used in this work is the ASR system called Audimus, which is described in (Meinedo et al., 2010). It is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of hidden Markov models with the pattern discriminative classification capabilities of multi-Layer perceptrons. The language and acoustic models were trained on newspaper texts and broadcast news (BN) speech data. The multiple-pronunciation EP lexicon includes about 114k entries. The Word Error Rate (WER) of our current ASR system is under 20% for BN speech in average: 18.4% obtained in one of our BN evaluation test sets (RTP07), composed by six one hour long news shows from 2007 (Meinedo et al., 2010).

In total, about 4.5K words were transcribed for the 11 interactions that were recorded so far, corresponding to 40 minutes of speech after removing silent segments. The average WER was 62%, with a smallest and largest WERs of 52% and 87% respectively. The very high WER value corresponds to what is usually expected when transcribing spontaneous speech. Many ASR errors were made on verb forms that were not in the BN lexicon. This is due to the fact that the speakers used verbs conjugated at the first person, which is not often observed in BN data. Some of these verb forms were not in the lexicon and the out-of-vocabulary rate was high, with a 2.2% value. In previous experiments, we obtained significant performance gains on elderly read speech by adapting the acoustic models on a subset. Nevertheless, no gain was obtained so far on the spontaneous small corpus.

5. Conclusions

In this paper, we presented a Wizard-Of-Woz interface from which the global architecture and the dialog control manager will be available on our Metashare Web page[3]. This interface was used to collect elderly speech in the context of a project on acoustic modeling of elderly speech in European Portuguese. The speech corpus along with manual orthographic transcriptions will also be available. Small adaptations were necessary to adapt the interface to elderly speakers. In particular, one cannot ask an elderly speaker to click on a "speech-to-talk" button before speaking. Despite their age, speakers above 75 years old were very enthusiastic about participating to the recordings and they seemed to like the avatar and the interface. Our automatic speech recognition baseline system was used to measure a WER of about 62%. Although this is a very high WER, its value corresponds to the performance observed when transcribing spontaneous speech in the literature, with also the increased difficulty to transcribe elderly speech.

6. Acknowledgment

This work was partially supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project PTDC/EEA-PLP/121111/2010 and under project PEst-OE/EEI/LA0021/2011.

[3] http://metanet4u.l2f.inesc-id.pt/

7. References

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., and Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan*, 87:7:49–57.

Dahlbaeck, N., Joensson, A., and Ahrenberg, L. (1993). Wizard of Oz studieswhy and how. *Knowledge-Based Systems*, 6:258–266.

Meinedo, H., Abad, A., Pellegrini, T., Neto, J., and Trancoso, I. (2010). The L2F Broadcast News Speech Recognition System. In *Proc. Fala*, pages 93–96, Vigo.

Paulo, S., Oliveira, L.-C., Mendes, C., Figueira, L., Cassaca, R., Viana, C., and Moniz, H. (2008). DIXI – A Generic Text-to-Speech System for European Portuguese. In *Computational Processing of the Portuguese Language*, volume 5190 of *LNAI*, pages 91–100. Springer-Verlag.

Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M., and Braga, D. (2012). Impact of age in ASR for the elderly: preliminary experiments in European Portuguese. In *Proc. IberSPEECH*, Madrid.

Vipperla, R., Renals, S., and Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. In *Proc. Interspeech*, pages 2550–2553, Brisbane.